

It's a Small World for Random Surfers

Abbas Mehrabian

Department of Combinatorics and Optimization, University of Waterloo

`amehrabi@uwaterloo.ca`

Nick Wormald*

School of Mathematical Sciences, Monash University

`nick.wormald@monash.edu`

Abstract

We prove logarithmic upper bounds for the diameters of the random-surfer Webgraph model and the PageRank-based selection Webgraph model, confirming the small world phenomenon holds for them. In the special case when the generated graph is a tree, we provide close lower and upper bounds for the diameters of both models.

Keywords: random-surfer Webgraph model, PageRank-based selection model, small-world phenomenon, height of random trees, probabilistic analysis, large deviations

1 Introduction

Due to the ever growing interest in social networks, the Webgraph, biological networks, etc., in recent years a great deal of research has been built around modelling real world networks (see, e.g., the monographs [6, 8, 10, 15]). One of the important observations about many real world networks involves the diameter, which is the maximum shortest-path distance between any two nodes. The so-called *small world phenomenon* is that the diameter of a network is significantly smaller than its size, typically growing as a polylogarithmic function.

The Webgraph is a directed graph whose vertices are the static web pages, and there is an edge joining two vertices if there is a hyperlink in the first page pointing to the second

*Supported by Australian Laureate Fellowships grant FL120100125.

page. Barabási and Albert [1] in 1999 introduced one of the first models for the Webgraph, widely known as the *preferential attachment model*. Their model can be informally described as follows (see [5] for the formal definition). Let d be a positive integer. We start with a fixed small graph, and in each time-step a new vertex appears and is joined to d old vertices, where the probability of joining to each old vertex is proportional to its *degree*. Pandurangan, Raghavan and Upfal [19] in 2002 introduced the *PageRank-based selection model* for the Webgraph. This model is similar to the previous model, except the attachment probabilities are proportional to the *PageRanks* of the vertices rather than their degrees. Blum, Chan, and Rwebangira [4] in 2006 introduced a *random-surfer model* for the Webgraph, in which the d out-neighbours of the new vertex are chosen by doing d independent random walks that start from random vertices and whose lengths are geometric random variables with parameter p . It was shown that under certain conditions, the previous two models are equivalent. See Section 2 for the formal definitions of these models, and the condition for their equivalence.

The directed models considered here generate directed acyclic graphs (new vertices create edges to old vertices), so it is natural to define the *diameter* of a directed graph as the maximum shortest-path distance between any two vertices in its underlying undirected graph. The diameter of the preferential attachment model was analysed by Bollobás and Riordan [5]. Previous work on the PageRank-based selection and random-surfer models has focused on their degree distributions. To the best of our knowledge, the diameters of these models have not been studied previously, and it is an open question even whether they have logarithmic diameter. One of the main contributions of this paper is giving logarithmic upper bounds for their diameters. We also give close lower and upper bounds in the special case $d = 1$, namely when the generated graph is (almost) a tree. It turns out that the key parameter in this case is the *height* of the generated random tree. We find the asymptotic value of the height for all $p \in [0.21, 1]$, and for $p \in (0, 0.21)$ we provide logarithmic lower and upper bounds. Our results hold *asymptotically almost surely (a.a.s.)*, which means the probability that they are true approaches 1 as the number of vertices grows.

1.1 Our approach and organization of the paper

In the preferential attachment model and most of its variations (see, e.g., [1, 13, 14, 17]) the probability that the new vertex attaches to an old vertex v , called the *attraction* of v , is proportional to a deterministic function of the degree of v . In other variations (see, e.g., [3, 16]) the attraction also depends on the so-called ‘fitness’ of v , which is a random variable generated independently for each vertex and does not depend on the structure of the graph. For analysing such models when they generate trees, a typical technique is to

approximate them with population-dependent branching processes and prove that results on the corresponding branching processes carry over to the original models. A classical example is Pittel [21] who estimated the height of random recursive trees. Bhamidi [2] used this technique to show that the height of a variety of preferential attachment trees is asymptotic to a constant times the logarithm of the number of vertices, where the constant depends on the parameters of the model.

In the random-surfer Webgraph model, however, the attraction of a vertex does not depend only on its degree, but rather on the graph's general structure, so the branching processes techniques cannot apply directly, and new ideas are needed.

The crucial novel idea in our proof is to reduce the attachment rule to a simple one, with the help of introducing (possibly negative) ‘weights’ for the edges. First, consider the general case, $d \geq 1$. Whenever a new vertex appears, it builds d new edges to old vertices; suppose that we mark the first new edge. Then the marked edges induce a spanning tree whose diameter we bound, and thus we get an upper bound for the diameter of the random-surfer Webgraph model.

In the special case $d = 1$, we obtain a *random recursive tree* with edge weights, and then we adapt a powerful technique developed by Broutin and Devroye [7] (that uses branching processes) to study its weighted height. This technique is based on large deviations. Their main theorem [7, Theorem 1] is not applicable here for two reasons. Firstly, the weights of edges on the path from the root to each vertex are not independent, and secondly, the weights can be negative.

We define the models and state our main results in Section 2. In Section 3 we give logarithmic upper bounds for the diameters of the random-surfer Webgraph model and the PageRank-based selection Webgraph model in the general case $d \geq 1$. In Sections 4–6 we focus on the special case $d = 1$ and prove close lower and upper bounds for the heights and diameters of the models. Section 4 contains the main technical contribution of this paper, where we explain how to transform the random-surfer tree model into one that is easier to analyse. The lower and upper bounds are proved in Sections 5 and 6, respectively. Concluding remarks appear in Section 7. For easing the flow of reading the paper, proofs of some technical lemmas has been put in the appendix.

2 Definitions and main results

Given $p \in (0, 1]$, let $\text{Geo}(p)$ denote a geometric random variable with parameter p ; namely for every nonnegative integer k , $\mathbb{P}[\text{Geo}(p) = k] = (1 - p)^k p$.

Definition (Random-Surfer Webgraph model [4]). Let d be a positive integer and let $p \in (0, 1]$. Generate a random directed rooted n -vertex multigraph, with all vertices having out-degree d . Start with a single vertex v_0 , the root, with d self-loops. At each subsequent step s , where $1 \leq s \leq n - 1$, a new vertex v_s appears and d edges are created from it to vertices in $\{v_0, v_1, \dots, v_{s-1}\}$, by doing the following probabilistic procedure d times, independently: choose a vertex u uniformly at random from $\{v_0, v_1, \dots, v_{s-1}\}$, and a fresh random variable $X = \text{Geo}(p)$; perform a simple random walk of length X starting from u , and join v_s to the last vertex of the walk.

The motivation behind this definition is as follows. Think of the vertex v_s as a new web page that is being set up. Say the owner wants to put d links in her web page. To build each link, she does the following: she goes to a random page. With probability p she likes the page and puts a link to that page. Otherwise, she clicks on a random link on that page, and follows the link to a new page. Again, with probability p she likes the new page and puts a link to that, otherwise clicks on a random link etc., until she finds a desirable page to link to. The geometric random variables correspond to this selection process.

Our main result regarding the diameter of the random-surfer Webgraph model is the following theorem (recall that the diameter of a directed graph is defined as the diameter of its underlying undirected graph). All logarithms are natural in this paper.

Theorem 1. *Let d be a positive integer and let $p \in (0, 1]$. A.a.s. as $n \rightarrow \infty$ the diameter of the random-surfer Webgraph model with parameters p and d is at most $8e^p(\log n)/p$.*

Notice that the upper bound in Theorem 1 does not depend on d (whereas one would expect that the diameter must decrease asymptotically as d increases). This independence is because in our argument we employ only the first edge created by each new vertex to bound the diameter.

When $d = 1$, we show in Theorem 3 below that the diameter is a.a.s. $\Theta(\log n)$. An interesting open problem is to evaluate the asymptotic value of the diameter when $d > 1$. In this regime the diameter might be of a smaller order, e.g. $\Theta(\log n / \log \log n)$, as is the case for the preferential attachment model (see [5, Theorem 1]).

A *random-surfer tree* is an undirected tree obtained from a random-surfer Webgraph with $d = 1$ by deleting the self-loops of the root and ignoring the edge directions. The *height*

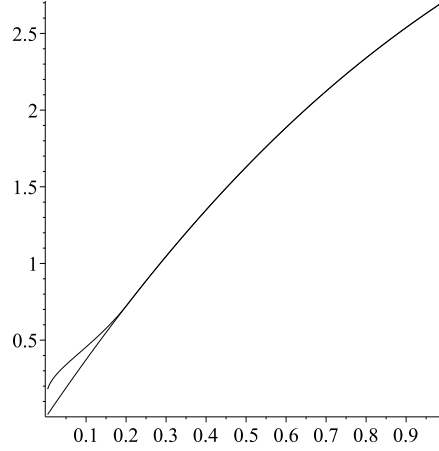


Figure 1: The functions c_L and c_U in Theorems 2 and 3.

of a tree is defined as the maximum graph distance between a vertex and the root. Our main result regarding the height of the random-surfer tree model is the following theorem.

Theorem 2. *For $p \in (0, 1)$, let $s = s(p)$ be the unique solution in $(0, 1)$ to*

$$s \log \left(\frac{(1-p)(2-s)}{1-s} \right) = 1. \quad (1)$$

Let $p_0 \approx 0.206$ be the unique solution in $(0, 1/2)$ to

$$\log \left(\frac{1-p}{p} \right) = \frac{1-p}{1-2p}. \quad (2)$$

Define the functions $c_L, c_U : (0, 1) \rightarrow \mathbb{R}$ as

$$c_L(p) = \exp(1/s)s(2-s)p,$$

and

$$c_U(p) = \begin{cases} c_L(p) & \text{if } p_0 \leq p < 1 \\ \left(\log \left(\frac{1-p}{p} \right) \right)^{-1} & \text{if } 0 < p < p_0. \end{cases}$$

For every fixed $\varepsilon > 0$, a.a.s. as $n \rightarrow \infty$ the height of the random-surfer tree model with parameter p is between $(c_L(p) - \varepsilon) \log n$ and $(c_U(p) + \varepsilon) \log n$.

The value p_0 and the functions c_L and c_U (plotted in Figure 1) are well defined by Lemma 16 below. Also, c_L and c_U are continuous, and $\lim_{p \rightarrow 0} c_L(p) = \lim_{p \rightarrow 0} c_U(p) = 0$ and $\lim_{p \rightarrow 1} c_L(p) = e$. We suspect that the gap between our bounds when $p < p_0$ is an artefact of

our proof technique, and we do not expect a phase transition in the behaviour of the height at $p = p_0$.

We also prove lower and upper bounds for the diameter, which are close to being tight.

Theorem 3. *Let c_L and c_U be defined as in Theorem 2. For every fixed $\varepsilon > 0$, a.a.s. as $n \rightarrow \infty$ the diameter of the random-surfer tree model with parameter $p \in (0, 1)$ is between $(2c_L(p) - \varepsilon) \log n$ and $(2c_U(p) + \varepsilon) \log n$.*

Immediately, we have the following corollary.

Corollary 4. *Let c_L and p_0 be defined as in Theorem 2. For any $p \in [p_0, 1)$, the height of the random-surfer tree model with parameter p is a.a.s. asymptotic to $c_L(p) \log n$ as $n \rightarrow \infty$, and its diameter is a.a.s. asymptotic to $2c_L(p) \log n$.*

A natural open problem is to close the gap between the lower and upper bounds in Theorems 2 and 3 when $p < p_0$. It seems that for solving this problem new ideas are required.

We now define the PageRank-based selection model introduced in [19, 20].

Definition (PageRank and the PageRank-based selection Webgraph model [19, 20]). Let d be a positive integer and let $p, \beta \in [0, 1]$. The *PageRank* of a directed graph is a probability distribution over its vertices, which is the stationary distribution of the following random walk. The random walk starts from a vertex chosen uniformly at random. In each step, with probability p it jumps to a vertex chosen uniformly at random, and with probability $1 - p$ it walks to a random out-neighbour of the current vertex.

The PageRank-based selection Webgraph model is a random n -vertex directed multigraph with all vertices having out-degree d , generated as follows. It starts with a single vertex with d self-loops. At each subsequent step a new vertex appears, chooses d old vertices and attaches to them (where a vertex can be chosen multiple times). These choices are independent and the head of each edge is a uniformly random vertex with probability β , and is a vertex chosen according to the PageRank distribution with probability $1 - \beta$.

The motivation behind this definition is as follows. Consider the case $\beta = 0$. Think of the vertex v_s as a new web page that is being set up. Say the owner wants to put d links in her web page. She finds the destination pages using d independent Google searches. Since Google sorts the search results according to their PageRank (see [18]), the probability that a given page is linked to is close to its PageRank.

Our main result regarding the diameter of the PageRank-based selection model is the following theorem.

Theorem 5. *Let d be a positive integer and let $p, \beta \in (0, 1]$. A.a.s. as $n \rightarrow \infty$ the diameter of the PageRank-based selection Webgraph model with parameters d , p , and β is at most $8e^p(\log n)/p$.*

Chebolu and Melsted [9, Theorem 1.1] showed the random-surfer Webgraph model is equivalent to the PageRank-based selection Webgraph model with $\beta = 0$ (this fact also follows from Lemma 9 in Section 3). Hence Theorems 1 follows immediately from Theorem 5. Moreover, the conclusions of Theorems 2 and 3 apply to the PageRank-based selection Webgraph model with $\beta = 0$ and $d = 1$.

In Theorems 2 and 3 we have assumed that $p < 1$, since the situation for $p = 1$ has been clarified in previous work. Let $p = 1$. Then a random-surfer tree has the same distribution as a so-called random recursive tree, the height of which is a.a.s. asymptotic to $e \log n$ as proved by Pittel [21]. It is not hard to alter the argument in [21] to prove that the diameter is a.a.s. asymptotic to $2e \log n$. The diameter of a random-surfer Webgraph thus has also an asymptotically almost sure upper bound of $2e \log n$. For the rest of the paper, we fix $p \in (0, 1)$.

We include some definitions here. Define the *depth* of a vertex as the length of a shortest path (ignoring edge directions) connecting the vertex to the root, and the *height* of a graph G , denoted by $\text{ht}(G)$, as the maximum depth of its vertices. Clearly the diameter is at most twice the height. In a weighted tree (a tree whose *edges* are weighted), define the *weight* of a vertex to be the sum of the weights of the edges connecting the vertex to the root, and the *weighted height* of tree T , written $\text{wht}(T)$, to be the maximum weight of its vertices. We view an unweighted tree as a weighted tree with unit edge weights, in which case the weight of a vertex is its depth, and the notion of weighted height is the same as the usual height.

We will need two large deviation inequalities, whose proofs are standard and can be found in the appendix.

Define the function $\Upsilon : (0, \infty) \rightarrow \mathbb{R}$ as

$$\Upsilon(x) = \begin{cases} x - 1 - \log(x) & \text{if } 0 < x \leq 1 \\ 0 & \text{if } 1 < x. \end{cases} \quad (3)$$

Lemma 6. *Let E_1, E_2, \dots, E_m be independent exponential random variables with mean 1. For any fixed $x > 0$, as $m \rightarrow \infty$ we have*

$$\exp(-\Upsilon(x)m - o(m)) \leq \mathbb{P}[E_1 + E_2 + \dots + E_m \leq xm] \leq \exp(-\Upsilon(x)m).$$

Define the function $f : (-\infty, 1] \rightarrow \mathbb{R}$ as

$$f(x) = (2 - x)^{2-x} p (1 - p)^{1-x} (1 - x)^{x-1}. \quad (4)$$

Lemma 7. *Let Z_1, Z_2, \dots, Z_m be independent $1 + \text{Geo}(p)$ random variables, and let $\kappa \geq 1/p$. Then we have $\mathbb{P}[Z_1 + Z_2 + \dots + Z_m \geq \kappa m] \leq f(2 - \kappa)^m$.*

3 Upper bound for the PageRank-based model

In this section we prove Theorem 5, which gives an upper bound for the diameter of the PageRank-based selection Webgraph model. Theorem 1 follows immediately using [9, Theorem 1.1]. We need a technical lemma, whose proof can be found in the appendix.

Lemma 8. *Let η, c be positive numbers satisfying $\eta \geq 4e^p/p$ and $c \leq p\eta$. Then we have $-c\Upsilon(1/c) + c \log f(2 - \eta/c) < \max\{\eta(1 - p) \log(1 - p^3), -0.15p\eta\} - 1$.*

We now describe an alternative way to generate the edge destinations in the PageRank-based selection model. Define the non-negative random variable \mathcal{L} as

$$\mathcal{L} = \mathcal{L}(p, \beta) = \begin{cases} 0 & \text{with probability } \beta, \\ \text{Geo}(p) & \text{with probability } 1 - \beta. \end{cases}$$

Note that $\text{Geo}(p)$ stochastically dominates \mathcal{L} .

Lemma 9. *The head of each new edge in the PageRank-based selection model can be obtained by sampling a vertex u uniformly from the existing graph and performing a simple random walk of length \mathcal{L} starting from u .*

The proof is a straightforward generalization of that of [9, Theorem 1.1].

Proof. Let G denote the existing graph, and let $\pi : V(G) \rightarrow [0, 1]$ denote the PageRank distribution. Then by definition, π is the unique probability distribution satisfying

$$\pi(v) = \frac{p}{|V(G)|} + (1 - p) \sum_{u \in V(G)} \frac{\pi(u) \cdot \#(uv)}{\text{out-deg}(u)}. \quad (5)$$

Here $\#(uv)$ denotes the number of copies of the directed edge uv in the graph (which is zero if there is no edge from u to v), and $\text{out-deg}(u)$ denotes the out-degree of u .

It suffices to show that if we sample a vertex uniformly and perform a random walk of length $\text{Geo}(p)$, the last vertex of the walk has distribution π . Let $\tau : V(G) \rightarrow [0, 1]$ denote

the probability distribution of the last vertex, let \mathcal{P} denote the probability transition matrix of the simple random walk, and let $\sigma = [1/|V(G)|, 1/|V(G)|, \dots, 1/|V(G)|]^T$ be the uniform distribution. Then we have

$$\tau = \sum_{k=0}^{\infty} (1-p)^k p \mathcal{P}^k \sigma = p\sigma + (1-p)\mathcal{P} \left(\sum_{k=1}^{\infty} (1-p)^{k-1} p \mathcal{P}^{k-1} \sigma \right) = p\sigma + (1-p)\mathcal{P}\tau.$$

Comparing with (5) and noting that the stationary distribution of an ergodic Markov chain is unique, we find that $\tau = \pi$, as required. \blacksquare

We now have the ingredients to prove Theorem 5.

Proof of Theorem 5. Let $\eta = 4e^p/p$. We define an auxiliary tree whose node set equals the vertex set of the graph generated by the PageRank-based selection Webgraph model, and whose weighted height dominates the height of this graph. Then we show a.a.s. this tree has weighted height at most $\eta \log n$, which completes the proof.

Initially the tree has just one vertex v_0 . By Lemma 9, the growth of the PageRank-based selection model at each subsequent step $s \in \{1, 2, \dots, n-1\}$ can be described as follows: a new vertex v_s appears and d edges are created from it to vertices in $\{v_0, v_1, \dots, v_{s-1}\}$, by doing the following probabilistic procedure d times, independently: choose a vertex u uniformly at random from $\{v_0, v_1, \dots, v_{s-1}\}$, and a fresh random variable \mathcal{L} ; perform a simple random walk of length \mathcal{L} starting from u , and join v_s to the last vertex of the walk.

Consider a step s and the first chosen $u \in \{v_0, \dots, v_{s-1}\}$ and \mathcal{L} . In the tree, we join the vertex v_s to u and set the weight of the edge $v_s u$ to be $\mathcal{L} + 1$. Note that the edge weights are mutually independent. Clearly, the weight of v_s in the auxiliary tree is greater than or equal to the depth of v_s in the graph. Hence, it suffices to show that a.a.s. the weighted height of the auxiliary tree is at most $\eta \log n$. We work with the tree in the rest of the proof.

Let us consider an alternative way to grow the tree, used by Devroye, Fawzi, and Fraiman [12], which results in the same distribution. Let U_1, U_2, \dots be i.i.d. uniform random variables in $(0, 1)$. Then for each new vertex v_s , we attach it to the vertex $v_{\lfloor s U_s \rfloor}$, which is indeed a vertex uniformly chosen from $\{v_0, \dots, v_{s-1}\}$.

For convenience, we consider the tree when it has $n+1$ vertices v_0, v_1, \dots, v_n . Let $D(s), W(s)$ denote the depth and the weight of vertex v_s , respectively. We have

$$\begin{aligned} \mathbb{P}[\text{wht}(\text{auxiliary tree}) > \eta \log n] &\leq \sum_{s=1}^n \mathbb{P}[W(s) > \eta \log n] \\ &\leq n \mathbb{P}[W(n) > \eta \log n] = \sum_{d=1}^n \mathcal{A}(d), \end{aligned}$$

where we define

$$\mathcal{A}(d) = n\mathbb{P}[D(n) = d] \mathbb{P}[W(n) > \eta \log n | D(n) = d] .$$

To complete the proof it is enough to show $\sum_{d=1}^n \mathcal{A}(d) = o(1)$.

Let $P(0) = 0$ and for $s = 1, \dots, n$, let $P(s)$ denote the index of the parent of v_s . We have

$$\mathbb{P}[D(n) \geq d] = \mathbb{P}[D(P(n)) \geq d-1] = \dots = \mathbb{P}[D(P^{d-1}(n)) \geq 1] = \mathbb{P}[P^{d-1}(n) \geq 1] .$$

Since $P(m) = \lfloor mU_m \rfloor \leq mU_m$ for each $0 \leq m \leq n$ and since the U_i are i.i.d., we have

$$\mathbb{P}[P^{d-1}(n) \geq 1] \leq \mathbb{P}[nU_1U_2 \dots U_{d-1} \geq 1] .$$

Let $E_i = -\log U_i$. Then E_i is exponential with mean 1, and moreover,

$$\begin{aligned} \mathbb{P}[D(n) \geq d] &\leq \mathbb{P}[nU_1U_2 \dots U_{d-1} \geq 1] \\ &= \mathbb{P}[E_1 + \dots + E_{d-1} \leq \log n] \leq \exp\left(-(d-1)\Upsilon\left(\frac{\log n}{d-1}\right)\right) , \end{aligned} \quad (6)$$

where we have used Lemma 6. The right-hand side is $o(1/n)$ for $d = 1.1e \log n$. Hence to complete the proof we need only show that

$$\mathcal{A}(d) = o(1/\log n) \quad \forall d \in (0, 1.1e \log n) . \quad (7)$$

Fix an arbitrary positive integer $d \in (0, 1.1e \log n)$. The random variable $W(n)$, conditional on $D(n) = d$, is a sum of d i.i.d. $1 + \mathcal{L}$ random variables. Since $\text{Geo}(p)$ stochastically dominates \mathcal{L} , by Lemma 7 and since $\eta > 1.1e/p$, we have

$$\mathbb{P}[W(n) > \eta \log n | D(n) = d] \leq f(2 - \eta \log n/d)^d , \quad (8)$$

where f is defined in (4).

Combining (6) and (8), we get

$$\mathcal{A}(d) \leq \exp\left[\log n - (d-1)\Upsilon\left(\frac{\log n}{d-1}\right) + d \log f(2 - \eta \log n/d)\right] . \quad (9)$$

Let $c = d/\log n$ and $c_1 = c - 1/\log n$. Let $\vartheta = \max\{\eta(1-p)\log(1-p^3), -0.15p\eta\}$. Note that ϑ is a negative constant. By Lemma 8 and since the function $c\Upsilon(1/c)$ is uniformly continuous on $[0, 1.1e]$, we find that for large enough n ,

$$-c_1\Upsilon(1/c_1) + c \log f(2 - \eta/c) < \vartheta/2 - 1 .$$

Together with (9), this gives $\mathcal{A}(d) \leq \exp(\vartheta \log n/2)$, and (7) follows. ■

4 Transformations of the random-surfer tree model

In Sections 4–6 we study the random-surfer tree model. In this section we show how to transform the random-surfer tree model three times to eventually obtain a new random tree model, which we analyse in subsequent sections. The first transformation is novel. The second one was perhaps first used by Broutin and Devroye [7], and the third one probably by Pittel [21].

Let us call the random-surfer tree model the *first model*. First, we will replace the attachment rule with a simpler one by introducing *weights* for the edges. In the first model, the edges are unweighted and in every step s a new vertex v_s appears, chooses an old vertex u , and attaches to a vertex in the path connecting u to the root, according to some rule. We introduce a second model that is weighted, and such that there is a one to one correspondence between the vertices in the second model and in the first model. For a vertex v in the first model, we denote its corresponding vertex in the second model by \bar{v} . In the second model, in every step s a new vertex \bar{v}_s appears, chooses an old vertex \bar{u} and attaches to \bar{u} , and the weight $w(\bar{u}\bar{v}_s)$ of the new edge $\bar{u}\bar{v}_s$ is chosen such that the *weight* of \bar{v}_s equals the *depth* of v_s in the first model. Let $w(\bar{u})$ denote the weight of vertex \bar{u} . Then it follows from the definition of the random-surfer tree model that $w(\bar{u}\bar{v}_s)$ is distributed as $\max\{1 - \text{Geo}(p), 1 - w(\bar{u})\}$. The term $1 - w(\bar{u})$ appears here solely because the weight of \bar{v}_s is at least 1 (in the first model, the depth of v_s is at least 1, since it cannot attach to a vertex higher than the root). Because the depth of v in the first model equals the weight of \bar{v} in the second model, the height of the first model equals the weighted height of the second model.

We will need to make the degrees of the tree bounded, so we define a third model. In this model, the new vertex can attach just to the leaves. In step s a new vertex v_s appears, chooses a random leaf u and joins to u using an edge with weight distributed as $\max\{1 - \text{Geo}(p), 1 - w(\bar{u})\}$. Simultaneously, a new vertex u' appears and joins to u using an edge with weight 0. Then we have $w(u) = w(u')$ and henceforth u' plays the role of u , i.e. the next vertex wanting to attach to u , but cannot do so because u is no longer a leaf, may attach to u' instead. Clearly there exists a coupling between the second and third models in which the weighted height of the third model, when it has $2n - 1$ vertices, equals the weighted height of the second model with n vertices. In fact the second model may be obtained from the third one by contracting all zero-weight edges. We can thus study the weighted height of the first model by studying it in the third model.

All the above models were defined using discrete time steps. We now define a fourth model using the following continuous time branching process, which we call \mathcal{P} . At time

0 the root is born. From this moment onwards, whenever a new vertex v is born (say at time κ), it waits for a random time E , which is distributed exponentially with mean 1, and after time E has passed (namely, at absolute time $\kappa + E$) gives birth to two children v_1 and v_2 , and dies. The weights of the edges vv_1 and vv_2 are generated as follows: vertex v chooses $i \in \{1, 2\}$ independently and uniformly at random. The weight of vv_i is distributed as $\max\{1 - \text{Geo}(p), 1 - w(v)\}$ and the weight of vv_{3-i} is 0. Given $t \geq 0$, we denote by T_t the almost surely finite random tree obtained by taking a snapshot of this process at time t . By the memorylessness of the exponential distribution, if one starts looking at this process at any deterministic moment, the next leaf to die is chosen uniformly at random. Hence for any stopping time τ , the distribution of T_τ , conditional on T_τ having $2n - 1$ vertices, is the same as the distribution of the third model when it has $2n - 1$ vertices.

The following lemma implies that certain results for T_t carry over to results for the random-surfer tree model.

Lemma 10. *Assume that there exist constants θ_L, θ_U such that for every fixed $\varepsilon > 0$,*

$$\mathbb{P}[\theta_L(1 - \varepsilon)t \leq \text{wht}(T_t) \leq \theta_U(1 + \varepsilon)t] \rightarrow 1$$

as $t \rightarrow \infty$. Then for every fixed $\varepsilon > 0$, a.a.s. as $n \rightarrow \infty$ the height of the random-surfer tree model is between $\theta_L(1 - \varepsilon) \log n$ and $\theta_U(1 + \varepsilon) \log n$.

Proof. Let $\ell_n = 2n - 1$, and let $\varepsilon > 0$ be fixed. For the process \mathcal{P} , we define three stopping times as follows:

a_1 is the deterministic time $(1 - \varepsilon) \log(\ell_n)$.

A_2 is the random time when the evolving tree has exactly ℓ_n vertices.

a_3 is the deterministic time $(1 + \varepsilon) \log(\ell_n)$.

By hypothesis, a.a.s. as $n \rightarrow \infty$ we have

$$(1 - \varepsilon)\theta_L \log(\ell_n) \leq \text{wht}(T_{a_1}) \text{ and } \text{wht}(T_{a_3}) \leq (1 + \varepsilon)\theta_U \log(\ell_n). \quad (10)$$

Broutin and Devroye [7, Proposition 2] considered the infinite process T_t as $t \rightarrow \infty$ and proved that almost surely

$$\lim_{t \rightarrow \infty} \frac{\log |V(T_t)|}{t} = 1,$$

which implies that a.a.s. as $t \rightarrow \infty$, we have $\log |V(T_t)| \sim t$. This means that, as $n \rightarrow \infty$, a.a.s.

$$\log |V(T_{a_1})| \sim a_1 = (1 - \varepsilon) \log(\ell_n),$$

and hence $|V(T_{a_1})| < \ell_n$, which implies $a_1 < A_2$. Symmetrically, it can be proved that a.a.s. as $n \rightarrow \infty$ we have $A_2 < a_3$. It follows that a.a.s. as $n \rightarrow \infty$

$$\text{wht}(T_{a_1}) \leq \text{wht}(T_{A_2}) \leq \text{wht}(T_{a_3}) . \quad (11)$$

On the other hand, as noted above, T_{A_2} has the same distribution as the third model with $2n-1$ vertices, whose weighted height has the same distribution as that of the random-surfer tree model with n vertices. Chaining (10) and (11) completes the proof. \blacksquare

It will be convenient to define T_t in a static way, which is equivalent to the dynamic definition above.

Definition (T_∞, T_t) . Let T_∞ denote an infinite binary tree. To every edge e is associated a random vector (E_e, W_e) and to every vertex v a random variable W_v , where the W_e 's and W_v 's are the *weights*. The law for $\{E_e\}_{e \in E(T)}$ is easy: first with every vertex v we associate independently an exponential random variable with mean 1, and we let the values of E on the edges joining v to its two children be equal to this variable. In the dynamic interpretation, this random variable denotes the length of life of v . Generation of the weights is done in a top-down manner, where we think of the root as the top vertex. Let the weight of the root be zero. Let v be a vertex whose weight has been determined, and let v_1, v_2 be its two children. Choose $i \in \{1, 2\}$ independently and uniformly at random, and then choose $Y = 1 - \text{Geo}(p)$ independently of previous choices. Then let

$$W_{vv_i} = \max\{Y, 1 - W_v\}, \quad W_{v_i} = W_v + W_{vv_i} , \quad (12)$$

and

$$W_{vv_j} = 0, \quad W_{v_j} = W_v$$

for $j = 3 - i$.

For a vertex v , let $\pi(v)$ be the set of edges of the unique path connecting v to the root. It is easy to check that the weight of any vertex v equals $\sum_{e \in \pi(v)} W_e$. We define the *birth time* of a vertex v , written B_v , as

$$B_v = \sum_{e \in \pi(v)} E_e ,$$

where the birth time of the root is defined as zero. Finally, given $t \geq 0$ we define T_t as the subtree of T_∞ induced by vertices with birth time at most t . Note that T_t is finite almost surely.

5 Lower bounds for the random-surfer tree model

Here we prove the lower bounds in Theorems 2 and 3. For this, we consider another infinite binary tree T'_∞ which is very similar to T_∞ , except for the generation rules for the weights, which are as follows. Let the weight of the root be zero. Let v be a vertex whose weight has been determined, and let v_1, v_2 be its two children. Choose $i \in \{1, 2\}$ independently and uniformly at random, and choose $Y = 1 - \text{Geo}(p)$ independently of previous choices. Then let

$$W_{vv_i} = Y \text{ and } W_{v_i} = W_v + W_{vv_i} \quad (13)$$

and

$$W_{vv_j} = 0 \text{ and } W_{v_j} = W_v$$

for $j = 3 - i$. Comparing (13) with (12), we find that the weight of every vertex in T'_∞ is stochastically less than or equal to that of its corresponding vertex in T_∞ . The tree T'_t is defined as before. Clearly probabilistic lower bounds for $\text{wht}(T'_t)$ are also probabilistic lower bounds for $\text{wht}(T_t)$. Distinct vertices u and v in a tree are called *antipodal* if the unique (u, v) -path in the tree passes through the root.

Lemma 11. *Consider the tree T'_∞ . Let $\gamma_L : (0, 1) \rightarrow \mathbb{R}$ be such that for every $a \in (0, 1)$, each vertex u and each descendent v of u that is m levels deeper,*

$$\mathbb{P}[W_v - W_u \geq am] \geq \exp(-m\gamma_L(a) - o(m)) \quad (14)$$

as $m \rightarrow \infty$. Assume that there exist $\alpha^, \rho^* \in (0, 1)$ with*

$$\gamma_L(\alpha^*) + \Upsilon(\rho^*) = \log 2. \quad (15)$$

Then for every fixed $\varepsilon > 0$, a.a.s. there exist antipodal vertices u, v of T'_t with weights at least $\frac{\alpha^}{\rho^*}(1 - \varepsilon)t$.*

The proof is very similar to the proof of [7, Lemma 4] except a small twist is needed at the end to handle the negative weights.

Proof. Let $c = \frac{\alpha^*}{\rho^*}$, and let $\varepsilon, \delta > 0$ be arbitrary. We prove that with probability at least $1 - \delta$ for all large enough t there exists a pair (u, v) of antipodal vertices of T'_∞ with $\max\{B_u, B_v\} < t$ and $\min\{W_u, W_v\} > (1 - 2\varepsilon)ct$.

Let L be a constant positive integer that will be determined later, and let $\alpha = \alpha^*$ and $\rho = \frac{\alpha}{c(1-\varepsilon)} > \rho^*$. By (15) and since $\rho^* < 1$ and Υ is strictly decreasing on $(0, 1]$, we have

$$\gamma_L(\alpha) + \Upsilon(\rho) < \log 2.$$

Build a Galton-Watson process from T'_∞ whose particles are a subset of vertices of T'_∞ , as follows. Start with the root as the initial particle of the process. If a given vertex u is a particle of the process, then its potential offspring are its 2^L descendants that are L levels deeper. Moreover, such a descendent v is an offspring of u if and only if $W_v - W_u \geq \alpha L$ and $B_v - B_u \leq \rho L$. As these two events are independent, the expected number of children of u is at least

$$2^L \mathbb{P}[W_v - W_u \geq \alpha L] \mathbb{P}[B_v - B_u \leq \rho L] \geq \exp[(\log 2 - \gamma_L(\alpha) - \Upsilon(\rho) - o(1))L]$$

as $L \rightarrow \infty$, by (14) and Lemma 6. Since we have $\log 2 - \gamma_L(\alpha) - \Upsilon(\rho) > 0$, we may choose L large enough that this expected value is strictly greater than 1. Therefore, this Galton-Watson process survives with probability $q > 0$.

We now boost this probability up to $1 - \delta$, by starting several independent processes, giving more chance that at least one of them survives. Specifically, let b be a constant large enough that

$$(1 - q)^{2^{b-1}} < \delta/3.$$

Consider 2^b Galton-Watson processes, which have the vertices at depth b of T'_∞ as their initial particles, and reproduce using the same rule as before. Let a be a constant large enough that

$$2^{b+1}(e^{-a} + (1 - p)^{a+2}) < \delta/3,$$

and let A be the event that all edges e in the top b levels of T'_∞ have $E_e \leq a$ and $W_e \geq -a$. Then

$$1 - \mathbb{P}[A] \leq 2^{b+1}(e^{-a} + (1 - p)^{a+2}) < \delta/3.$$

Also, let Q be the event that in each of the two branches of the root, at least one of the 2^{b-1} Galton-Watson processes survives. Then

$$1 - \mathbb{P}[Q] \leq 2(1 - q)^{2^{b-1}} < 2\delta/3,$$

and so with probability at least $1 - \delta$ both A and Q occur.

Assume that both A and Q occur. Let

$$m = \left\lfloor \frac{t(1 - \varepsilon)}{\rho L} \right\rfloor$$

and let u and v be particles at generation m of surviving processes in distinct branches of the root. Then u and v are antipodal,

$$\max\{B_u, B_v\} \leq ab + m\rho L \leq t(1 - \varepsilon) + O(1) < t,$$

and

$$\min\{W_u, W_v\} \geq -ab + m\alpha L \geq \frac{(1-\varepsilon)\alpha}{\rho} t - O(1) > c(1-2\varepsilon)t$$

for t large enough, as required. ■

Let Y_1, Y_2, \dots be i.i.d. with $Y_i = 1 - \text{Geo}(p)$. Recall the definition of $f : (-\infty, 1] \rightarrow \mathbb{R}$ from (4):

$$f(x) = (2-x)^{2-x} p (1-p)^{1-x} (1-x)^{x-1}.$$

Note that $f(1) = p$ since by convention $0^0 = 1$, and $f(2-p^{-1}) = 1$. The following lemma follows by noting that f is positive and the derivative of $\log f$ is $\log\left(\frac{1-x}{(2-x)(1-p)}\right)$.

Lemma 12. *The function f is continuous in $(-\infty, 1]$ and differentiable in $(-\infty, 1)$. Moreover, f is increasing on $(-\infty, 2-p^{-1}]$ and decreasing on $[2-p^{-1}, 1]$.*

Lemma 13. (a) *There is an absolute constant C such that for any $a \in [2-p^{-1}, 1]$ and any positive integer m we have*

$$\mathbb{P}[Y_1 + \dots + Y_m \geq am] \leq C m f(a)^m.$$

(b) *As $m \rightarrow \infty$, uniformly for all $a \in [0, 1]$ we have*

$$\mathbb{P}[Y_1 + \dots + Y_m \geq am] \geq [f(a) - o(1)]^m.$$

(c) *If $p \geq 1/2$, then as $m \rightarrow \infty$, uniformly for all $a \in [0, 2 - \frac{1}{p}]$ we have*

$$\mathbb{P}[Y_1 + \dots + Y_m \geq am] \geq [1 - o(1)]^m.$$

Proof. The conclusions are easy to see for $a = 1$, so assume that $a < 1$. First, assume that am is an integer. Consider a sequence of independent biased coin flips, each of which is heads with probability p . A random walker starts from 0, takes one step to the right on seeing heads, and one to the left on seeing tails. Then $Y_1 + \dots + Y_m$ is the walker's position just after seeing the m -th head. Thus $Y_1 + \dots + Y_m = am$ if and only if the $(2m - am)$ -th coin comes up heads, and in the first $2m - am$ coin flips we see exactly m heads and $m - am$ tails, so we have

$$\begin{aligned} \mathbb{P}[Y_1 + \dots + Y_m = am] &= \binom{2m - am - 1}{m - 1} p^m (1-p)^{m-am} \\ &= \Theta\left(\binom{2m - am}{m} p^m (1-p)^{m-am}\right) \\ &= \Theta(f(a)^m / \sqrt{m}), \end{aligned} \tag{16}$$

where we have used Stirling's approximation for the last equality.

(a) Let $a \in [2 - p^{-1}, 1)$, and let C be an absolute constant for the upper bound of Θ in (16). Then

$$\begin{aligned}\mathbb{P}[Y_1 + \dots + Y_m \geq am] &\leq m \sup\{\mathbb{P}[Y_1 + \dots + Y_m = \alpha m] : \alpha \in [a, 1]\} \\ &\leq C\sqrt{m} [\sup\{f(\alpha) : \alpha \in [a, 1]\}]^m \leq Cm(f(a))^m\end{aligned}$$

since f is decreasing on $[2 - \frac{1}{p}, 1]$ by Lemma 12 and m is a positive integer.

(b) Assume that $m \rightarrow \infty$. Then

$$\mathbb{P}[Y_1 + \dots + Y_m \geq am] \geq \mathbb{P}[Y_1 + \dots + Y_m = \lceil am \rceil] = (f(a) - o(1))^m$$

uniformly for all $a \in [0, 1)$ by continuity of f .

(c) Assume that $p \geq 1/2$ and that $m \rightarrow \infty$. Then

$$\begin{aligned}\mathbb{P}[Y_1 + \dots + Y_m \geq am] &\geq \mathbb{P}[Y_1 + \dots + Y_m = \lceil (2 - p^{-1})m \rceil] \\ &= (f(2 - p^{-1}) - o(1))^m = (1 - o(1))^m\end{aligned}$$

uniformly for all $a \in [0, 2 - \frac{1}{p}]$ by continuity of f and since $f(2 - p^{-1}) = 1$. ■

We define a two variable function

$$\Phi(a, s) = p(1 - p)(2 - s)^2(s - a) - a(1 - s), \quad (17)$$

and we define a function $\phi : [0, 1] \rightarrow [0, 1]$ as follows: given $a \in [0, 1]$, $\phi(a)$ is the unique solution in $[a, 1]$ to

$$\Phi(a, \phi(a)) = p(1 - p)(2 - \phi(a))^2(\phi(a) - a) - a(1 - \phi(a)) = 0. \quad (18)$$

Lemma 14(a) below shows that ϕ is well defined. The proof of this lemma is straightforward and can be found in the appendix.

Lemma 14. (a) Given $a \in [0, 1]$, there is a unique solution $s \in [a, 1]$ to $\Phi(a, s) = 0$. If $a \in \{0, 1\}$ then $\phi(a) = a$. If $a \in (0, 1)$ then $0 < a < \phi(a) < 1$.

(b) If $s = \phi(a)$ then

$$\frac{sf(s)^{a/s}}{a^{a/s}(s - a)^{1 - \frac{a}{s}}} = \frac{s}{s - a} \left(\frac{1 - s}{(1 - p)(2 - s)} \right)^a.$$

(c) The function ϕ is increasing on $[0, 1]$ and differentiable on $(0, 1)$.

(d) The function ϕ is invertible and ϕ^{-1} is increasing. If $s \in \{0, 1\}$ then $\phi^{-1}(s) = s$. If $s \in (0, 1)$ then $0 < \phi^{-1}(s) < s < 1$.

Next let $\hat{Y}_1, \hat{Y}_2, \dots$ be independent and distributed as follows: for every $i = 1, 2, \dots$ we flip an unbiased coin, if it comes up heads, then $\hat{Y}_i = Y_i$, otherwise $\hat{Y}_i = 0$.

Define the function $g_L : (0, 1) \rightarrow \mathbb{R}$ as

$$g_L(a) = \begin{cases} 1/2 & \text{if } p > 1/2 \text{ and } 0 < a < 1 - \frac{1}{2p} \\ \frac{\phi(a)-a}{\phi(a)} \left(\frac{(1-p)(2-\phi(a))}{1-\phi(a)} \right)^a & \text{otherwise.} \end{cases}$$

Note that g_L is continuous as $\phi(1 - \frac{1}{2p}) = 2 - 1/p$. The proofs of the following two lemmas are standard and can be found in the appendix.

Lemma 15. *We have the following large deviation inequality for every fixed $a \in (0, 1)$ as $m \rightarrow \infty$.*

$$\mathbb{P} \left[\hat{Y}_1 + \dots + \hat{Y}_m \geq am \right] \geq (2g_L(a) - o(1))^{-m}.$$

Lemma 16. (a) *There exists a unique solution $p_0 \in (0, 1/2)$ to*

$$\log \left(\frac{1-p}{p} \right) = \frac{1-p}{1-2p}.$$

Also, if $p \leq p_0$ then $\log \left(\frac{1-p}{p} \right) \geq \frac{1-p}{1-2p}$.

(b) *Given $p \in (0, 1)$, there exists a unique solution $s_0 \in (0, 1)$ to*

$$(1-p)(2-s) = \exp(1/s)(1-s).$$

Moreover, if $p > 1/2$ then $s_0 > 2 - p^{-1}$, and if $p_0 < p \leq 1/2$ then $s_0 > \frac{1-2p}{1-p}$.

Lemma 17. *Given $\varepsilon > 0$, a.a.s as $t \rightarrow \infty$ there exist two antipodal vertices u, v of T'_t with weights at least $c_L(p)(1-\varepsilon)t$. In particular, a.a.s. the weighted height of T'_t is at least $c_L(p)(1-\varepsilon)t$.*

Proof. By Lemma 16(b), there is a unique solution $s \in (0, 1)$ to

$$(1-p)(2-s) = \exp(1/s)(1-s).$$

By the definition of c_L ,

$$c_L = c_L(p) = \exp(1/s)s(2-s)p.$$

Lemma 15 implies that the assumption (14) of Lemma 11 holds for $\gamma_L(a) = \log(2g_L(a))$. Let $a = \phi^{-1}(s)$ and let $\rho = 1 - \frac{a}{s}$. Since $s \in (0, 1)$ we have $0 < a < s < 1$ by Lemma 14(d), and thus $\rho \in (0, 1)$ as well. Moreover, since $\Phi(a, s) = 0$, we have $c_L = a/\rho$.

We now show that $g_L(a) = \frac{s-a}{s} \exp(a/s)$. This is clear if $p \leq 1/2$, so assume that $p > 1/2$. It is easy to verify that $\Phi(1 - \frac{1}{2p}, 2 - \frac{1}{p}) = 0$. Since $p > 1/2$, by Lemma 16(b) we have $s > 2 - \frac{1}{p}$. Since ϕ^{-1} is increasing, we have $a = \phi^{-1}(s) \geq 1 - \frac{1}{2p}$.

From $g_L(a) = \frac{s-a}{s} \exp(a/s)$ we get

$$\log(2g_L(a)) + \rho - 1 - \log(\rho) = \log 2,$$

and Lemma 11 completes the proof. ■

The lower bound in Theorem 2 follows easily from Lemmas 17 and 10.

Proof of the lower bound in Theorem 3. Fix $\varepsilon > 0$. Let us define the *semi-diameter* of a tree as the maximum weighted distance between any two antipodal vertices. Clearly, semi-diameter is a lower bound for the diameter, so we just need to show a.a.s. as $n \rightarrow \infty$ the semi-diameter of the random-surfer model with n vertices is at least $(2c_L(p) - \varepsilon) \log n$. By Lemma 17, a.a.s. as $t \rightarrow \infty$ the semi-diameter of T'_t is at least $(2c_L(p) - \varepsilon)t$. Using an argument similar to the proof of Lemma 10 we may conclude that a.a.s. as $n \rightarrow \infty$ the semi-diameter of the third model (of Section 4) with $2n - 1$ vertices is at least $(2c_L(p) - \varepsilon) \log n$. It is easy to observe that this statement is also true for the random-surfer model with n vertices, and the proof is complete. ■

6 Upper bounds for the random-surfer tree model

In this section we prove the upper bounds in Theorems 2 and 3.

Lemma 18. *Let $\gamma_U : [0, 1] \rightarrow [0, \infty)$ be a continuous function such that for every fixed $a \in [0, 1]$ and every vertex v of T_∞ at depth m ,*

$$\mathbb{P} \left[\sum_{e \in \pi(v)} W_e > am \right] \leq \exp(-m\gamma_U(a) + o(m)) \quad (19)$$

as $m \rightarrow \infty$. Define

$$\theta = \sup \left\{ \frac{a}{\rho} : \gamma_U(a) + \Upsilon(\rho) = \log 2 : a \in [0, 1], \rho \in (0, \infty) \right\}. \quad (20)$$

Then for every fixed $\varepsilon > 0$,

$$\mathbb{P}[\text{wht}(T_t) > \theta(1 + \varepsilon)t] \rightarrow 0$$

as $t \rightarrow \infty$.

The proof is similar to the proof of [7, Lemma 3], in which the assumption (19) is not needed. In fact, in the model studied in [7], the weights $\{W_e : e \in \pi(v)\}$ are mutually independent, and the authors use Cramér's Theorem to obtain a large deviation inequality for $\sum_{e \in \pi(v)} W_e$, which is similar to (19).

Proof. We first prove a claim.

Claim. For every $\varepsilon > 0$ there exists $\delta > 0$ such that for all $\rho \in \left(0, \frac{1}{\theta(1+\varepsilon)}\right]$,

$$\Upsilon(\rho) + \gamma_U(\theta(1+\varepsilon)\rho) - \log 2 \geq \delta.$$

Proof of Claim. Assume that this is not the case for some $\varepsilon > 0$. This means there exists a sequence $(\rho_i)_{i=1}^\infty$ such that for all $i = 1, 2, \dots$,

$$\Upsilon(\rho_i) + \gamma_U(\theta(1+\varepsilon)\rho_i) - \log 2 < 1/i.$$

Then $(\rho_i)_{i=1}^\infty$ has a convergent subsequence. Let $\rho^* \in \left[0, \frac{1}{\theta(1+\varepsilon)}\right]$ be the limit. It cannot be the case that $\rho^* = 0$ since $\Upsilon(x) \rightarrow \infty$ as $x \rightarrow 0$, and γ_U is non-negative. By continuity of Υ and γ_U we have

$$\Upsilon(\rho^*) + \gamma_U(\theta(1+\varepsilon)\rho^*) - \log 2 \leq 0.$$

Since Υ is continuous, decreasing, and attains all values in $[0, \infty)$, we can choose $\rho' \leq \rho^*$ so that

$$\Upsilon(\rho') + \gamma_U(\theta(1+\varepsilon)\rho^*) - \log 2 = 0,$$

But then

$$\frac{\theta(1+\varepsilon)\rho^*}{\rho'} \geq \theta(1+\varepsilon) > \theta,$$

contradicting the definition of θ in (20). □

Fix $\varepsilon > 0$ and let A_k be the event that there exists a vertex at depth k of T_t with weight larger than $\theta(1+\varepsilon)t$. By the union bound,

$$\mathbb{P}[\text{wht}(T_t) > \theta(1+\varepsilon)t] \leq \sum_{k=1}^{\infty} \mathbb{P}[A_k] = \sum_{k > \theta(1+\varepsilon)t} \mathbb{P}[A_k],$$

as the weights of all edges are at most 1.

Let $k > \theta(1+\varepsilon)t$. A vertex v at depth k of T_∞ is included in T_t and has weight larger than $\theta(1+\varepsilon)t$ if and only if $B_v \leq t$ and $W_v > \theta(1+\varepsilon)t$. These two events are independent

by the definition of T_∞ . The random variable B_v is distributed as a sum of k independent exponential random variables with mean 1, and so

$$\begin{aligned}\mathbb{P}[B_v \leq t, W_v > \theta(1 + \varepsilon)t] &\leq \exp \left[\left(-\Upsilon(t/k) - \gamma_U \left(\frac{\theta(1 + \varepsilon)t}{k} \right) + o(1) \right) k \right] \\ &\leq \exp [(-\log 2 - \delta + o(1))k] ,\end{aligned}$$

where we have used Lemma 6 and (19) for the first inequality, and $\delta > 0$ is the constant provided by the claim. Since there are 2^k vertices at depth k of T_∞ , by the union bound

$$\mathbb{P}[A_k] \leq 2^k \exp [(-\log 2 - \delta + o(1))k] \leq \exp [(-\delta + o(1))k] .$$

For t large enough the $o(1)$ term is less than $\delta/2$, and thus

$$\begin{aligned}\mathbb{P}[\text{wht}(T_t) > \theta(1 + \varepsilon)t] &\leq \sum_{k > \theta(1 + \varepsilon)t} \mathbb{P}[A_k] \leq \sum_{k > \theta(1 + \varepsilon)t} \exp [(-\delta/2)k] \\ &= O(\exp [(-\delta/2)\theta(1 + \varepsilon)t]) = o(1) .\end{aligned}$$

■

Let Y_1, Y_2, \dots be i.i.d. with $Y_i = 1 - \text{Geo}(p)$, and define random variables X_1, X_2, \dots as follows:

$$X_1 = \max\{Y_1, 1\} ,$$

and for $i \geq 1$,

$$X_{i+1} = \max\{Y_{i+1}, 1 - (X_1 + \dots + X_i)\} .$$

Define the function $h : [0, 1] \rightarrow \mathbb{R}$ as

$$h(x) = \begin{cases} 1 & \text{if } p \geq \frac{1}{2} \text{ and } 0 \leq x \leq 2 - \frac{1}{p} \\ \left(\frac{p}{1-p}\right)^x & \text{if } p < \frac{1}{2} \text{ and } 0 \leq x \leq \frac{1-2p}{1-p} \\ (2-x)^{2-x} p (1-p)^{1-x} (1-x)^{x-1} & \text{otherwise .} \end{cases} \quad (21)$$

Note that in the third case we have $h(x) = f(x)$, where f is defined in (4). It is easy to see that h is continuous. The proof of the following lemma can be found in the appendix.

Lemma 19. *There exists an absolute constant C such that for every $a \in [0, 1]$ and every positive integer m we have $\mathbb{P}[X_1 + \dots + X_m > am] \leq C m^2 h(a)^m$.*

Next we define random variables $\hat{X}_1, \hat{X}_2, \dots$ as follows: for every $i = 1, 2, \dots$ we flip an independent unbiased coin, if it comes up heads, then $\hat{X}_i = X_i$, otherwise $\hat{X}_i = 0$.

We define the function $g_U : [0, 1] \rightarrow \mathbb{R}$ as

$$g_U(a) = \begin{cases} 1/2 & \text{if } p \geq 1/2 \text{ and } 0 \leq a \leq 1 - 1/2p \\ (\frac{1-p}{p})^a / 2 & \text{if } p < 1/2 \text{ and } 0 \leq a \leq \frac{1-2p}{2-2p} \\ \frac{1}{p} & \text{if } a = 1 \\ \frac{\phi(a)-a}{\phi(a)} \left(\frac{(1-p)(2-\phi(a))}{1-\phi(a)} \right)^a & \text{otherwise,} \end{cases} \quad (22)$$

where ϕ is defined by (18). Note that by Lemma 14(a), we have $0 < a < \phi(a) < 1$ for $a \in (0, 1)$, so g_U is well defined for all $a \in [0, 1]$. The proof of the following lemma can be found in the appendix.

Lemma 20. (a) *We have the following large deviation inequality for every $a \in [0, 1]$ and every positive integer m , where C' is an absolute constant:*

$$\mathbb{P} \left[\hat{X}_1 + \dots + \hat{X}_m > am \right] \leq C' m^3 (2g_U(a))^{-m}. \quad (23)$$

(b) *The function g_U is continuously differentiable on $(0, 1)$ and*

$$g'_U(a) = \begin{cases} 0 & \text{if } p \geq 1/2 \text{ and } 0 < a \leq 1 - 1/2p \\ \log(\frac{1-p}{p}) g_U(a) & \text{if } p < 1/2 \text{ and } 0 < a \leq \frac{1-2p}{2-2p} \\ \log \left(\frac{(1-p)(2-\phi(a))}{1-\phi(a)} \right) g_U(a) & \text{otherwise.} \end{cases} \quad (24)$$

(c) *The function $\log g_U(a)$ is increasing and convex. It is strictly increasing when $g_U(a) > 1/2$.*

Lemma 21. *Let $\omega > 0$ and let $\tau(x) : [0, \omega] \rightarrow \mathbb{R}$ be a positive function that is differentiable on $(0, \omega)$ and satisfies*

$$\alpha(x) + \chi(\tau(x)) = 0 \quad \forall x \in [0, \omega] \quad (25)$$

for convex functions α, χ , with α increasing and χ decreasing. Assume there exists $x^ \in (0, \omega)$ such that $\tau'(x^*) = \tau(x^*)/x^*$. Then we have*

$$\frac{x^*}{\tau(x^*)} \geq \frac{y}{\tau(y)} \quad (26)$$

for all $y \in [0, \omega]$.

Proof. We first prove that τ is convex and increasing. Pick $x_1, x_2 \in [0, \omega]$ and $\lambda_1, \lambda_2 \in [0, 1]$ with $\lambda_1 + \lambda_2 = 1$. We need to show that

$$\tau(\lambda_1 x_1 + \lambda_2 x_2) \leq \lambda_1 \tau(x_1) + \lambda_2 \tau(x_2). \quad (27)$$

We have

$$\begin{aligned}
\chi(\lambda_1\tau(x_1) + \lambda_2\tau(x_2)) &\leq \lambda_1\chi(\tau(x_1)) + \lambda_2\chi(\tau(x_2)) \\
&= -\lambda_1\alpha(x_1) - \lambda_2\alpha(x_2) \\
&\leq -\alpha(\lambda_1x_1 + \lambda_2x_2) \\
&= \chi(\tau(\lambda_1x_1 + \lambda_2x_2))
\end{aligned}$$

by convexity of χ , then (25), then convexity of α , and then (25) again. The equation (27) follows since χ is decreasing. Hence τ is convex. Also, τ is increasing since α is increasing and χ is decreasing.

Now, let $y \in [0, \omega]$. We prove (26) for $y < x^*$. The proof for $y > x^*$ is similar. By the mean value theorem, there exists $z \in (y, x^*)$ with

$$\tau'(z) = \frac{\tau(x^*) - \tau(y)}{x^* - y}.$$

On the other hand, since $z < x^*$ and τ is convex, we have

$$\tau'(z) \leq \tau'(x^*) = \frac{\tau(x^*)}{x^*}.$$

The inequality (26) follows from these two results. ■

We are ready to prove the upper bound in Theorem 2. The upper bound in Theorem 3 follows immediately as in every tree the diameter is at most twice the height.

Proof of the upper bound in Theorem 2. Let $c_U = c_U(p)$. By Lemma 10 we just need to show that given $\varepsilon > 0$, a.a.s as $t \rightarrow \infty$ the weighted height of T_t is at most $(1 + \varepsilon)c_U t$. For proving this we use Lemma 18. Lemma 20 implies that condition (19) of Lemma 18 holds with $\gamma_U(a) = \log(2g_U(a))$, so we need only show that

$$c_U = \sup \left\{ \frac{a}{\rho} : \log(g_U(a)) + \rho - 1 - \log(\rho) = 0 : a \in [0, 1], \rho \in (0, \infty) \right\}. \quad (28)$$

The function $\rho - 1 - \log(\rho)$ attains all values in $[0, \infty)$ for $\rho \in (0, 1]$. Moreover, it is strictly decreasing for $\rho \in (0, 1]$ and equals 0 for $\rho \in [1, \infty)$. So $\log(g_U(a)) + \rho - 1 - \log(\rho) = 0$ has a unique solution (for ρ) if $0 < g_U(a) < 1$, and no solution if $g_U(a) > 1$. Since $g_U(0) = 1/2$ and $g_U(1) = 1/p$, and the function $g_U(x)$ is continuous and strictly increasing when $g_U(x) > 1/2$, there is a unique x with $g_U(x) = 1$. Denote this point by a_{\max} . Define the function $\tau : [0, a_{\max}] \rightarrow (0, 1]$ as follows. Let $\tau(a_{\max}) = 1$ and for $x < a_{\max}$ let $\tau(x)$ be the unique number satisfying

$$\log(g_U(x)) + \tau(x) - 1 - \log \tau(x) = 0. \quad (29)$$

Hence to prove (28) it is enough to show that

$$c_U = \sup \left\{ \frac{x}{\tau(x)} : x \in [0, a_{\max}] \right\}. \quad (30)$$

We prove (30) using Lemma 21. The function $\log(g_U(a))$ is increasing and convex by Lemma 20(c), and it is easy to check that the function $\rho - 1 - \log(\rho)$ is decreasing and convex. Moreover, differentiating (29) gives

$$\frac{g'_U(x)}{g_U(x)} + \tau'(x) - \frac{\tau'(x)}{\tau(x)} = 0.$$

So by the implicit function theorem τ is differentiable in $x \in (0, a_{\max})$ and

$$\tau'(x) = \frac{\tau(x)}{1 - \tau(x)} \frac{g'_U(x)}{g_U(x)}.$$

By Lemma 21, we just need to show the existence of $x^* \in (0, a_{\max})$ with

$$c_U = \frac{x^*}{\tau(x^*)} = \frac{1 - \tau(x^*)}{\tau(x^*)} \frac{g_U(x^*)}{g'_U(x^*)}. \quad (31)$$

We consider two cases. Recall that $p_0 \approx 0.206$ is the solution to

$$\log \left(\frac{1-p}{p} \right) = \frac{1-p}{1-2p},$$

which has a unique solution by Lemma 16(a).

Case 1: $0 < p \leq p_0$. In this case we have

$$c_U = \left(\log \left(\frac{1-p}{p} \right) \right)^{-1}.$$

Let

$$a^* = \left[2 \log \left(\frac{1-p}{p} \right) \right]^{-1}.$$

By Lemma 16(a)

$$\log \left(\frac{1-p}{p} \right) \geq \frac{1-p}{1-2p},$$

which gives $a^* \leq \frac{1-2p}{2-2p}$, thus

$$g_U(a^*) = \left(\frac{1-p}{p} \right)^{a^*} / 2 = \exp \left(\frac{1}{2} - \log 2 \right) < 1$$

by the definition of g_U in (22), and

$$g'_U(a^*) = \log \left(\frac{1-p}{p} \right) g_U(a^*)$$

by Lemma 20(b). The definition of τ in (29) implies $\tau(a^*) = 1/2$. Moreover,

$$\frac{a^*}{\tau(a^*)} = \left(\log \left(\frac{1-p}{p} \right) \right)^{-1} = \frac{1 - \tau(a^*)}{\tau(a^*)} \frac{g_U(a^*)}{g'_U(a^*)},$$

which gives (31). Finally, since $g_U(a^*) < 1$, we have $a^* \in (0, a_{\max})$, and the proof is complete.

Case 2: $p_0 < p < 1$. In this case we have

$$c_U = ps^*(2 - s^*) \exp(1/s^*),$$

where $s^* \in (0, 1)$ is the unique solution for

$$s^* \log \left(\frac{(1-p)(2-s^*)}{1-s^*} \right) = 1. \quad (32)$$

Lemma 16(b) implies that s^* is well defined. Let $a^* = \phi^{-1}(s^*)$.

We first show that

$$g_U(a^*) = \frac{s^* - a^*}{s^*} \left(\frac{(1-p)(2-s^*)}{1-s^*} \right)^{a^*}. \quad (33)$$

If $p > 1/2$, then by Lemma 16(b) we have $s^* > 2 - \frac{1}{p}$. It is easy to verify that $\Phi(1 - \frac{1}{2p}, 2 - \frac{1}{p}) = 0$. Since ϕ^{-1} is increasing, we have $a^* = \phi^{-1}(s^*) > 1 - \frac{1}{2p}$, so (33) agrees with the definition of g_U in (22).

If $p_0 < p \leq 1/2$, then by Lemma 16(b) we have $s^* > \frac{1-2p}{1-p}$. It is easy to verify that $\Phi(\frac{1-2p}{2-2p}, \frac{1-2p}{1-p}) = 0$. Since ϕ^{-1} is increasing, we have $a^* = \phi^{-1}(s^*) > \frac{1-2p}{2-2p}$, so (33) agrees with the definition of g_U in (22).

Using (32), the equation (33) simplifies into

$$g_U(a^*) = \left(1 - \frac{a^*}{s^*} \right) \exp(a^*/s^*) < \exp(-\frac{a^*}{s^*}) \exp(a^*/s^*) = 1, \quad (34)$$

and by Lemma 20(b) we have

$$g'_U(a^*) = g_U(a^*) \log \left(\frac{(1-p)(2-s^*)}{1-s^*} \right) = g_U(a^*)/s^*.$$

It follows from (34) and the definition of τ in (29) that $\tau(a^*) = 1 - \frac{a^*}{s^*}$. Using (32) and $\Phi(a^*, s^*) = 0$, we get

$$\frac{a^*}{\tau(a^*)} = ps^*(2 - s^*) \exp(1/s^*) = \frac{1 - \tau(a^*)}{\tau(a^*)} \frac{g_U(a^*)}{g'_U(a^*)},$$

which gives (31). Finally, since $g_U(a^*) < 1$, we have $a^* \in (0, a_{\max})$, and the proof is complete. ■

7 Concluding Remarks

There is a common generalization of random recursive trees, preferential attachment trees, and random-surfer trees. Consider i.i.d. random variables $X_1, X_2, \dots \in \{0, 1, 2, \dots\}$. Start with a single vertex v_0 . At each step s a new vertex v_s appears, chooses a random vertex u in the present graph, and then walks X_s steps from u towards v_0 , joining to the last vertex in the walk (if it reaches v_0 before X_s steps, it joins to v_0). Random recursive trees correspond to $X_i = 0$, preferential attachment trees correspond to $X_i = \text{Bernoulli}(1/2)$ (see, e.g., [4, Theorem 3.1]), and random-surfer trees correspond to $X_i = \text{Geo}(p)$. Using the ideas of this paper, it is possible to obtain lower and upper bounds for the height and the diameter of this general model (similar to Theorems 2 and 3), provided one can prove large deviation inequalities (similar to Lemma 13) for the sum of X_i 's and also large deviation inequalities (similar to Lemma 19) for the sum of random variables X'_i , defined as

$$X'_1 = 1, \quad X'_{i+1} = \max\{1 - X_i, 1 - (X'_1 + \dots + X'_i)\}.$$

References

- [1] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.
- [2] S. Bhamidi. Universal techniques to analyze preferential attachment trees: global and local analysis. preprint, available via <http://www.unc.edu/~bhamidi/>, 2007.
- [3] G. Bianconi and A.-L. Barabási. Competition and multiscaling in evolving networks. *Europhys. Lett.*, 54(4):436–442, 2001.
- [4] A. Blum, T.-H. H. Chan, and M. R. Rwebangira. A random-surfer web-graph model. In *Proc. of 8th Workshop on Algorithm Engineering and Experiments and 3rd Workshop on Analytic Algorithmics and Combinatorics*, pages 238–246, 2006.
- [5] B. Bollobás and O. Riordan. The diameter of a scale-free random graph. *Combinatorica*, 24(1):5–34, January 2004.
- [6] A. Bonato. *A course on the web graph*, volume 89 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2008.
- [7] N. Broutin and L. Devroye. Large deviations for the weighted height of an extended class of trees. *Algorithmica*, 46(3-4):271–297, 2006.

- [8] D. Chakrabarti and C. Faloutsos. *Graph Mining: Laws, Tools, and Case Studies*. Synthesis Lectures on Data Mining and Knowledge Discovery. Morgan & Claypool Publishers, 2012.
- [9] P. Chebolu and P. Melsted. Pagerank and the random surfer model. In *Proceedings of the 19th annual ACM-SIAM symposium on Discrete algorithms*, SODA '08, pages 1010–1018, Philadelphia, PA, USA, 2008.
- [10] F. Chung and L. Lu. *Complex graphs and networks*, volume 107 of *CBMS Regional Conference Series in Mathematics*. Published for the Conference Board of the Mathematical Sciences, Washington, DC, 2006.
- [11] A. Dembo and O. Zeitouni. *Large deviations techniques and applications*, volume 38 of *Stochastic Modelling and Applied Probability*. Springer-Verlag, Berlin, 2010. Corrected reprint of the second (1998) edition.
- [12] L. Devroye, O. Fawzi, and N. Fraiman. Depth properties of scaled attachment random recursive trees. *Random Structures Algorithms*, 41(1):66–98, 2012.
- [13] S. Dommers, R. van der Hofstad, and G. Hooghiemstra. Diameters in preferential attachment models. *Journal of Statistical Physics*, 139(1):72–107, 2010.
- [14] E. Drinea, A. Frieze, and M. Mitzenmacher. Balls and bins models with feedback. In *Proceedings of the thirteenth annual ACM-SIAM symposium on Discrete algorithms*, SODA '02, pages 308–315, Philadelphia, PA, USA, 2002.
- [15] R. Durrett. *Random graph dynamics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 2010.
- [16] G. Ergün and G.J. Rodgers. Growing random networks with fitness. *Physica A: Statistical Mechanics and its Applications*, 303(1–2):261–272, 2002.
- [17] P. L. Krapivsky and S. Redner. Organization of growing random networks. *Phys. Rev. E*, 63:066123, May 2001.
- [18] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, 1999.
- [19] G. Pandurangan, P. Raghavan, and E. Upfal. Using pagerank to characterize web structure. In *Proceedings of the 8th Annual International Conference on Computing and Combinatorics*, COCOON '02, pages 330–339, London, UK, UK, 2002.

- [20] G. Pandurangan, P. Raghavan, and E. Upfal. Using pagerank to characterize web structure. *Internet Mathematics*, 3(1):1–20, 2006.
- [21] B. Pittel. Note on the heights of random recursive trees and random m -ary search trees. *Random Structures and Algorithms*, 5(2):337–347, 1994.

Appendix: omitted proofs

Proof of Lemma 6. We first prove the upper bound. If $x > 1$ then $\exp(-\Upsilon(x)m) = 1$, so we may assume that $0 < x \leq 1$. We use Chernoff’s technique. Let $\theta = 1 - 1/x$. Then we have

$$\begin{aligned} \mathbb{P}[E_1 + E_2 + \cdots + E_m \leq xm] &= \mathbb{P}[\exp(\theta E_1 + \cdots + \theta E_m) \geq \exp(\theta xm)] \\ &\leq \mathbb{E}[\exp(\theta E_1 + \cdots + \theta E_m)] / \exp(\theta xm) \\ &= \mathbb{E}[\exp(\theta E_1)] \mathbb{E}[\exp(\theta E_2)] \cdots \mathbb{E}[\exp(\theta E_m)] / \exp(\theta xm) \\ &= (1 - \theta)^{-m} \exp(-\theta xm) = \exp(-\Upsilon(x)m). \end{aligned}$$

We now prove the lower bound. If $x > 1$, then the result follows from Markov’s inequality, so we may assume that $0 < x \leq 1$. Let $\Lambda^*(x) = \sup\{\lambda x - \log(\mathbb{E}[e^{\lambda E_1}]) : \lambda \leq 0\}$. Since $\mathbb{E}[e^{\lambda E_1}] = 1/(1 - \lambda)$ for all $\lambda < 1$, the supremum here occurs at $\lambda = 1 - 1/x$, which implies $\Lambda^*(x) = \Upsilon(x)$. Then by Cramér’s Theorem (see, e.g., [11, Theorem 2.2.3, p. 27]) we have

$$\mathbb{P}[E_1 + E_2 + \cdots + E_m \leq xm] = \exp(-\Lambda^*(x)m + o(m)) = \exp(-\Upsilon(x)m + o(m)),$$

as required. ■

Proof of Lemma 7. We use Chernoff’s technique. Let θ satisfy

$$e^\theta = \frac{\kappa - 1}{\kappa(1 - p)}.$$

We have

$$\mathbb{E}[\exp(\theta Z_1)] = \sum_{k=1}^{\infty} p(1 - p)^{k-1} e^{\theta k} = \frac{pe^\theta}{1 - e^\theta(1 - p)}.$$

Thus we have

$$\begin{aligned} \mathbb{P}[Z_1 + Z_2 + \cdots + Z_m \geq \kappa m] &= \mathbb{P}[\exp(\theta Z_1 + \cdots + \theta Z_m) \geq \exp(\theta \kappa m)] \\ &\leq \mathbb{E}[\exp(\theta Z_1 + \cdots + \theta Z_m)] / \exp(\theta \kappa m) \\ &= \mathbb{E}[\exp(\theta Z_1)] \mathbb{E}[\exp(\theta Z_2)] \cdots \mathbb{E}[\exp(\theta Z_m)] / \exp(\theta \kappa m) \\ &= \left(\frac{pe^{\theta - \theta \kappa}}{1 - e^\theta(1 - p)} \right)^m = f(2 - \kappa)^m. \end{aligned} \quad \blacksquare$$

Proof of Lemma 8. We consider two cases.

Case 1: $c \geq 1$. In this case we prove

$$-c\Upsilon(1/c) + c \log f(2 - \eta/c) < \eta(1 - p) \log(1 - p^3) - 1.$$

Notice that we have $1 - c\Upsilon(1/c) = c + c \log(1/c)$, so, using the definition of f and since $\eta(1 - p) \leq \eta - c$, the conclusion is implied by

$$c + c \log(1/c) + \eta \log(\eta(1 - p)/(\eta - c)) + c \log(p(\eta - c)/((1 - p)c)) < (\eta - c) \log(1 - p^3).$$

Letting $r = \eta/c$ and since $c > 0$, this statement is equivalent to

$$ep(1 - p)^{r-1}r^2(r/(r - 1))^{r-1} < \eta(1 - p^3)^{r-1}.$$

Since $(r/(r - 1))^{r-1} < e$, and $1 - p < (1 - p^3)e^{-p}$, for this inequality to hold it suffices to have

$$e^{2+p}r^2p \exp(-pr) \leq 4e^p/p \quad \forall r \in [p^{-1}, \infty),$$

which follows from the fact that $x^2e^{-x} \leq 4e^{-2}$ for all $x \geq 1$.

Case 2: $c < 1$. In this case we prove

$$-c\Upsilon(1/c) + c \log f(2 - \eta/c) < -0.15p\eta - 1.$$

Since $\Upsilon(1/c) = 0$, this is equivalent to

$$1 + 0.15p\eta + c \log f(2 - \eta/c) < 0. \tag{35}$$

Note that

$$\left(\frac{\eta/c}{\eta/c - 1} \right)^{\eta/c-1} < e,$$

so we have

$$\begin{aligned} c \log f(2 - \eta/c) &= \log((\eta/c)^\eta p^c (1 - p)^{\eta-c} (\eta/c - 1)^{c-\eta}) \\ &< \log((e\eta p/c)^c (1 - p)^{\eta-c}) \leq c \log(e\eta p/c) + cp - p\eta, \end{aligned}$$

where we have used $\log(1 - p) \leq -p$ in the last inequality. Hence to prove (35), since $c > 0$, it suffices to show that

$$\frac{1}{c} + 1 + \log(\eta p/c) + p < 0.85p\eta/c. \tag{36}$$

Since $p\eta \geq 4e^p \geq 4 > 4c$, we have

$$\begin{aligned} \frac{1}{c} &< 0.25p\eta/c, \\ 1 + p &< \frac{1+p}{c} < \frac{e^p}{c} \leq 0.25p\eta/c, \\ \log(\eta p/c) &< 0.35p\eta/c, \end{aligned}$$

which imply (36). ■

Proof of Lemma 14. (a) The conclusion is clear for $a \in \{0, 1\}$, so we may assume that $a \in (0, 1)$. Since $\Phi(a, a) < 0$ and $\Phi(a, 1) > 0$, there exists at least one $s \in (a, 1)$ with $\Phi(a, s) = 0$. We now show that there is a unique such s . Fixing a , since Φ is differentiable with respect to s , it is enough to show that

$$\text{if } \Phi(a, s) = 0, \text{ then } \frac{\partial \Phi}{\partial s} > 0 \quad (37)$$

Let $\sigma = p(1 - p)$. We have

$$\frac{\partial \Phi}{\partial s} = \sigma(2 - s)^2 + a - 2\sigma(2 - s)(s - a).$$

At a point (a, s) with $\Phi(a, s) = 0$, we have

$$\sigma(2 - s)^2 = \frac{a(1 - s)}{s - a}, \text{ and } \sigma(2 - s)(s - a) = \frac{a(1 - s)}{2 - s},$$

so at this point,

$$\frac{\partial \Phi}{\partial s} = a \left(\frac{1}{s - a} + \frac{1}{1 - s} - \frac{2}{2 - s} \right),$$

which is strictly positive because

$$\min \left\{ \frac{1}{s - a}, \frac{1}{1 - s} \right\} > 1 > \frac{1}{2 - s},$$

and this proves (37).

(b) Plugging the definition of f from (4) and using $\Phi(a, s) = 0$ gives this equation.

(c) We first show that ϕ is differentiable and increasing on $(0, 1)$. Let $a \in (0, 1)$ and let $s = \phi(a)$. We have

$$\frac{\partial \Phi}{\partial a} = s - 1 - p(1 - p)(2 - s)^2 < 0,$$

and $\partial \Phi / \partial s$ is positive as proved in part (a). Hence by the implicit function theorem ds/da exists and is positive, so ϕ is differentiable and increasing on $(0, 1)$. Since $\phi(0) = 0$ and $\phi(1) = 1$, ϕ is increasing on $[0, 1]$.

(d) Let $s \in [0, 1]$. Then $\Phi(0, s)\Phi(s, s) \leq 0$ and so there exists at least one $a_0 \in [0, s]$ with $\Phi(a_0, s) = 0$. The function $\Phi(a, s)$ is linear in a and the coefficient of a is non-zero, hence this root a_0 is unique. The function ϕ^{-1} is increasing since ϕ is increasing. The last two statements follow from similar statements proved for ϕ in (a). ■

Proof of Lemma 15. We have

$$\mathbb{P} \left[\hat{Y}_1 + \cdots + \hat{Y}_m \geq am \right] = \sum_{k=\lceil am \rceil}^m \binom{m}{k} 2^{-m} \times \mathbb{P} [Y_1 + \cdots + Y_k \geq am],$$

where k denotes the number of \hat{Y}_i 's whose value was determined to be equal to Y_i .

If $p > 1/2$ and $0 < a < 1 - \frac{1}{2p}$, then letting $k = \lceil m/2 \rceil$ gives

$$\binom{m}{k} 2^{-m} = \Omega\left(\frac{1}{\sqrt{m}}\right)$$

by Stirling's approximation, and

$$\mathbb{P}[Y_1 + \dots + Y_k \geq am] \geq (1 - o(1))^k$$

by Lemma 13(c). This gives

$$\mathbb{P}[\hat{Y}_1 + \dots + \hat{Y}_m \geq am] \geq (1 - o(1))^m,$$

as required.

Otherwise, let $s = \phi(a)$. Then letting $k = \lceil am/s \rceil$ gives

$$\binom{m}{k} 2^{-m} = \Omega\left(\left[\frac{s(s-a)^{a/s}}{2(s-a)a^{a/s}}\right]^m\right) / m^2$$

by Stirling's approximation, and

$$\mathbb{P}[Y_1 + \dots + Y_k \geq am] \geq (f(s) - o(1))^k$$

by Lemma 13(b) and since f is continuous. Lemma 14(b) completes the proof. ■

Proof of Lemma 16. (a) The function $r(p) = \log\left(\frac{1-p}{p}\right) - \frac{1-p}{1-2p}$ approaches $+\infty$ when $p \rightarrow 0^+$ and approaches $-\infty$ when $p \rightarrow \frac{1}{2}^-$. Moreover,

$$r'(p) = \frac{-1}{p(1-p)} - \frac{1}{(1-2p)^2} < 0$$

for $p \in (0, 1/2)$. Hence $r(p)$ has a unique root p_0 , and $r(p) \geq 0$ if and only if $p \leq p_0$.

(b) The function

$$\mu(s) = \log(1-p) + \log(2-s) - \log(1-s) - \frac{1}{s}$$

approaches $-\infty$ as $s \rightarrow 0^+$, and approaches $+\infty$ as $s \rightarrow 1^-$, and its derivative is positive in $(0, 1)$, hence it has a unique root s_0 in $(0, 1)$. Also we have $\mu(2-p^{-1}) = p/(1-2p)$, which means that if $p > 1/2$ then $s_0 > 2-p^{-1}$. Moreover, if $p_0 < p \leq 1/2$, then by part (a),

$$\mu\left(\frac{1-2p}{1-p}\right) = \log\left(\frac{1-p}{p}\right) - \frac{1-p}{1-2p} = r(p) < 0,$$

which means $s_0 > \frac{1-2p}{1-p}$. ■

Proof of Lemma 19. The conclusion is obvious if $p \geq \frac{1}{2}$ and $a \leq 2 - \frac{1}{p}$, or if $a = 0$, since in these cases $h(a) = 1$. Also, $\mathbb{P}[X_1 + \dots + X_m > m] = 0$ so the conclusion is true if $a = 1$, so we may assume that $\max\{0, 2 - \frac{1}{p}\} < a < 1$.

Observe that if $X_1 + \dots + X_m > am$, there is a subsequence of the form Y_{m-k+1}, \dots, Y_m whose sum is at least am , and this subsequence contains at least am elements since $Y_i \leq 1$ for all i . Hence we have

$$\mathbb{P}[X_1 + \dots + X_m > am] \leq m \max\{\mathbb{P}[Y_1 + \dots + Y_k \geq am] : k \in [am, m] \cap \mathbb{Z}\}$$

as the Y_i 's are i.i.d.

For any integer $k \in [am, m]$, by Lemma 13(a) we have

$$\mathbb{P}[Y_1 + \dots + Y_k \geq am] \leq Ck(f(am/k))^k$$

for an absolute constant C , since $am/k \geq a > 2 - \frac{1}{p}$. Let $r = k/m \in [a, 1]$. So we find that

$$\mathbb{P}[X_1 + \dots + X_m > am] \leq Cm^2 \left(\sup\{f(a/r)^r : r \in [a, 1]\} \right)^m.$$

Let us define

$$\xi(r) = f(a/r)^r = (2r - a)^{2r-a} p^r (1-p)^{r-a} (r-a)^{a-r} r^{-r}.$$

So to complete the proof we just need to show that

$$\sup\{\xi(r) : r \in [a, 1]\} \leq h(a) \quad \forall a \in \left(\max\left\{0, 2 - \frac{1}{p}\right\}, 1 \right). \quad (38)$$

The function $\xi(r)$ is positive and differentiable for each $a \in (0, 1)$, hence the supremum here occurs either at a boundary point or at a point with zero derivative. The derivative of $\log(\xi(r))$ equals

$$\log \left(\frac{p(1-p)(2r-a)^2}{r(r-a)} \right).$$

Thus $\xi'(r)$ has the same sign as $\bar{\xi}(r) = p(1-p)(2r-a)^2 - r(r-a)$ in $r \in [a, 1]$. Notice that $\bar{\xi}(r)$ has two roots

$$r_1 = \frac{ap}{2p-1}, \text{ and } r_2 = \frac{a(1-p)}{1-2p}.$$

We may consider several cases.

Case 0: $p = 1/2$. The function $\bar{\xi}$ is positive, so ξ is increasing in $[a, 1]$, hence the supremum in (38) happens at $r = 1$ and its value is $f(a)$.

Case 1: $p > 1/2$. Since $a > 2 - \frac{1}{p}$, we find that $r_1 > 1$ and $r_2 < 0$. Moreover, $\bar{\xi}(a) \geq 0$. Thus $\bar{\xi}$ is non-negative in $[a, 1]$, which implies ξ is increasing in $[a, 1]$. Thus the supremum in (38) happens at $r = 1$ and its value is $f(a)$.

Case 2: $p < 1/2$ and $a \leq \frac{1-2p}{1-p}$. In this case $r_1 < 0$ and $a \leq r_2 \leq 1$. Since $\bar{\xi}(a) \geq 0$ and $\bar{\xi}(r_1) = \bar{\xi}(r_2) = 0$ and $\bar{\xi}$ is quadratic, the function $\bar{\xi}$ goes from positive to negative at r_2 . Therefore, the function ξ attains its supremum at r_2 and the supremum value in (38) equals

$$\xi(r_2) = \left(\frac{p}{1-p} \right)^a.$$

Case 3: $p < 1/2$ and $a > \frac{1-2p}{1-p}$. We find that $r_1 < 0$ and $r_2 > 1$, and $\bar{\xi}(a) \geq 0$, so $\bar{\xi}$ is non-negative in $[a, 1]$, hence ξ is increasing in $[a, 1]$. Thus the supremum in (38) happens at $r = 1$ and its value is $f(a)$. This completes the proof of (38) and the lemma. \blacksquare

Proof of Lemma 20. (a) First, the case $a = 0$ is obvious since $g_U(0) = 1/2$, and the case $a = 1$ is easy since $\mathbb{P}[\hat{X}_1 + \dots + \hat{X}_m > m] = 0$. So we may assume that $a \in (0, 1)$.

Letting k of the \hat{X}_i 's being equal to X_i and the rest equal to zero, we get

$$\begin{aligned} \mathbb{P}[\hat{X}_1 + \dots + \hat{X}_m > am] &= \sum_{k=am}^m \binom{m}{k} 2^{-m} \mathbb{P}[X_1 + \dots + X_k > am] \\ &\leq m \sup \left\{ \binom{m}{rm} 2^{-m} \mathbb{P}[X_1 + \dots + X_{rm} > am] : r \in [a, 1] \right\}. \end{aligned}$$

For a given $r \in [a, 1]$, Lemma 19 gives

$$\mathbb{P}[X_1 + \dots + X_{rm} > am] \leq C(rm)^2 h(a/r)^{rm} \leq Cm^2 h(a/r)^{rm}.$$

Moreover, by Stirling's approximation

$$\binom{m}{rm} = O\left(\frac{1}{r^{rm}(1-r)^{(1-r)m}}\right).$$

So, we find that

$$\mathbb{P}[\hat{X}_1 + \dots + \hat{X}_m > am] \leq C'm^3 \left[\sup \left\{ \frac{h(a/r)^r}{2r^r(1-r)^{1-r}} : r \in [a, 1] \right\} \right]^m.$$

Thus to complete the proof of part (a) we just need to show

$$g_U(a) = \inf \left\{ \frac{\zeta - a}{\zeta} \left(\frac{a}{(\zeta - a)h(\zeta)} \right)^{a/\zeta} : \zeta \in [a, 1] \right\}, \quad (39)$$

where we have used the change of variable $\zeta = a/r$. For analysing this infimum we define the two variable function

$$\psi(a, \zeta) = \frac{\zeta - a}{\zeta} \left(\frac{a}{(\zeta - a)h(\zeta)} \right)^{a/\zeta}$$

with domain $\{(a, \zeta) : 0 < a < 1, a \leq \zeta \leq 1\}$, and consider two cases depending on the value of p .

Case 1: $p \geq 1/2$. By the definition of h in (21) we have

$$\psi(a, \zeta) = \begin{cases} \frac{\zeta-a}{\zeta} \left(\frac{a}{\zeta-a} \right)^{a/\zeta} & \text{if } a \leq \zeta \leq 2 - p^{-1} \\ \frac{\zeta-a}{\zeta} \left(\frac{a}{(\zeta-a)f(\zeta)} \right)^{a/\zeta} & \text{otherwise,} \end{cases}$$

where f is defined in (4). Since $f(2 - p^{-1}) = 1$, ψ is continuous here. Let us define $\psi_1(\zeta) = \frac{\zeta-a}{\zeta} \left(\frac{a}{\zeta-a} \right)^{a/\zeta}$ and $\psi_2(\zeta) = \frac{\zeta-a}{\zeta} \left(\frac{a}{(\zeta-a)f(\zeta)} \right)^{a/\zeta}$.

The derivative of $\log \psi_1(\zeta)$ is

$$a \log \left(\frac{\zeta-a}{a} \right) / \zeta^2,$$

which is negative for $\zeta < 2a$ and positive for $\zeta > 2a$. This implies $\psi_1(\zeta)$ is decreasing when $\zeta \leq 2a$ and increasing when $\zeta \geq 2a$. So ψ_1 achieves its minimum at $\zeta = 2a$, and its minimum value is $1/2$.

The derivative of $\log \psi_2(\zeta)$ is

$$\frac{a}{\zeta^2} [\log(p(1-p)(2-\zeta)^2(\zeta-a)) - \log(a(1-\zeta))].$$

Comparing with (17) we find that this derivative has the same sign as $\Phi(a, \zeta)$. So by Lemma 14(a) it vanishes at a unique point $\zeta = \phi(a)$. Also at $\zeta = \phi(a)$ we have $\partial\Phi/\partial\zeta > 0$ (see (37)), which implies $\Phi(a, \zeta)$ is non-positive when $\zeta \leq \phi(a)$ and non-negative when $\zeta \geq \phi(a)$. Thus ψ_2 achieves its minimum at $\phi(a)$, and its minimum value is

$$\psi_2(\phi(a)) = \frac{\phi(a)-a}{\phi(a)} \left(\frac{a}{(\phi(a)-a)f(\phi(a))} \right)^{a/\phi(a)} = \frac{\phi(a)-a}{\phi(a)} \left(\frac{(1-p)(2-\phi(a))}{1-\phi(a)} \right)^a$$

by Lemma 14(b).

We conclude that:

(i) If $2a \leq 2 - 1/p$, then the infimum of ψ occurs at $\zeta = 2a$ and its value is $\psi(a, 2a) = \psi_1(2a) = 1/2$. The reason is that on $[a, 2 - 1/p]$, $\psi = \psi_1$ achieves its minimum at $2a$, and on $[2 - 1/p, 1]$, $\psi = \psi_2$ is increasing since $\Phi(a, 2 - 1/p) \geq 0$.

(ii) If $a \leq 2 - 1/p$ and $2a > 2 - 1/p$, then the infimum occurs at $\zeta = \phi(a)$ and its value is $\frac{\phi(a)-a}{\phi(a)} ((1-p)(2-\phi(a))/(1-\phi(a)))^a$. The reason is that on $[a, 2 - 1/p]$, $\psi = \psi_1$ is decreasing, and on $[2 - 1/p, 1]$, $\psi = \psi_2$ achieves its minimum at $\phi(a)$ since $\Phi(a, 2 - 1/p) \leq 0$ and $\Phi(a, 1) \geq 0$.

(iii) If $a > 2 - 1/p$, then the infimum occurs at $\zeta = \phi(a)$ and its value is equal to $\frac{\phi(a)-a}{\phi(a)} ((1-p)(2-\phi(a))/(1-\phi(a)))^a$. The reason is that on $[a, 1]$, $\psi = \psi_2$ achieves its minimum at $\phi(a)$ since $\Phi(a, a) \leq 0$ and $\Phi(a, 1) \geq 0$.

Case 2: $p < 1/2$. By the definition of h in (21) we have

$$\psi(a, \zeta) = \begin{cases} \left(\frac{1-p}{p}\right)^a \frac{\zeta-a}{\zeta} \left(\frac{a}{\zeta-a}\right)^{a/\zeta} & \text{if } a \leq \zeta \leq \frac{1-2p}{1-p} \\ \frac{\zeta-a}{\zeta} \left(\frac{a}{(\zeta-a)f(\zeta)}\right)^{a/\zeta} & \text{otherwise.} \end{cases}$$

The function ψ is continuous here since

$$f\left(\frac{1-2p}{1-p}\right) = \left(\frac{p}{1-p}\right)^{\frac{1-2p}{1-p}}.$$

Let us define $\psi_3(\zeta) = \left(\frac{1-p}{p}\right)^a \frac{\zeta-a}{\zeta} \left(\frac{a}{\zeta-a}\right)^{a/\zeta}$. Since $\psi_3(\zeta) = \left(\frac{1-p}{p}\right)^a \psi_1(\zeta)$, the function $\psi_3(\zeta)$ is decreasing when $\zeta \leq 2a$ and increasing when $\zeta \geq 2a$. So ψ_3 achieves its minimum at $\zeta = 2a$ and its minimum value is $\left(\frac{1-p}{p}\right)^a / 2$. We conclude that

(iv) If $a \leq 1 - p/(1-p)$ and $2a \leq 1 - p/(1-p)$, then the infimum in (39) occurs at $\zeta = 2a$ and at this point we have $\psi(a, \zeta) = \left(\frac{1-p}{p}\right)^a / 2$. The reason is that on $[a, 1 - p/(1-p)]$, $\psi = \psi_3$ achieves its minimum at $2a$, and on $[1 - p/(1-p), 1]$, $\psi = \psi_2$ is increasing since $\Phi(a, 1 - p/(1-p)) \geq 0$.

(v) If $a \leq 1 - p/(1-p)$ and $2a > 1 - p/(1-p)$, then the infimum in (39) occurs at $\zeta = \phi(a)$ and its value is equal to $\frac{\phi(a)-a}{\phi(a)} ((1-p)(2-\phi(a))/(1-\phi(a)))^a$. The reason is that on $[a, 1 - p/(1-p)]$, $\psi = \psi_3$ is decreasing, and on $[1 - p/(1-p), 1]$, $\psi = \psi_2$ achieves its minimum at $\phi(a)$ since $\Phi(a, 1 - p/(1-p)) \leq 0$ and $\Phi(a, 1) \geq 0$.

(vi) If $a > 1 - p/(1-p)$, then the infimum in (39) occurs at $\zeta = \phi(a)$ and its value is equal to $\frac{\phi(a)-a}{\phi(a)} ((1-p)(2-\phi(a))/(1-\phi(a)))^a$. The reason is that on $[a, 1]$, $\psi = \psi_2$ achieves its minimum at $\phi(a)$ since $\Phi(a, a) \leq 0$ and $\Phi(a, 1) \geq 0$.

In all cases we proved that $g_U(a)$ actually gives the value of the infimum in (39), and this concludes the proof of (39) and of part (a).

(b) Consider the definition of g_U in (22). The formulae in (24) for the cases ' $p \geq 1/2$ and $0 < a \leq 1 - 1/2p$ ' and ' $p < 1/2$ and $0 < a \leq \frac{1-2p}{2-2p}$ ' are clearly true, so we assume that a is in the 'otherwise' case. We use the equality (39). Note that as proved in part (a), the infimum

in (39) occurs at the point $\zeta = \phi(a)$ that has $\left. \frac{\partial \psi}{\partial \zeta} \right|_{(a, \phi(a))} = 0$. This implies for every a_0 ,

$$\begin{aligned} \frac{dg_U}{da}(a_0) &= \frac{\partial \psi}{\partial a}(a_0, \phi(a_0)) + \frac{\partial \psi}{\partial \zeta}(a_0, \phi(a_0)) \times \frac{d\phi}{da}(a_0) \\ &= \frac{\partial \psi}{\partial a}(a_0, \phi(a_0)) = \frac{\partial}{\partial a} \left[\frac{\zeta - a}{\zeta} \left(\frac{a}{(\zeta - a)f(\zeta)} \right)^{a/\zeta} \right] \Big|_{(a_0, \phi(a_0))}, \end{aligned}$$

and (24) follows from computing this partial derivative and putting $\zeta = \phi(a_0)$.

We next prove the continuity of g_U and its derivative. Note that by Lemma 14(a), if $a \in (0, 1)$ then $\phi(a) \in (0, 1)$. First, g_U is continuous at $a = 1$ since

$$\begin{aligned} \lim_{a \rightarrow 1} \frac{\phi(a) - a}{\phi(a)} \left(\frac{(1-p)(2-\phi(a))}{1-\phi(a)} \right)^a &= \lim_{a \rightarrow 1} \frac{\phi(a) - a}{\phi(a)} \left(\frac{a}{p(2-\phi(a))(\phi(a)-a)} \right)^a \\ &= \lim_{a \rightarrow 1} \frac{(\phi(a)-a)^{1-a}}{\phi(a)} \left(\frac{a}{p(2-\phi(a))} \right)^a = \frac{1}{p}. \end{aligned}$$

For $p \geq 1/2$, the only discontinuity for g_U can possibly occur at $b = 1 - 1/2p$. However at this point we have $\phi(b) = 2b = 2 - p^{-1}$ so that $(1-p)(2-\phi(b)) = 1 - \phi(b)$. Hence the left and right limits of g_U equal $1/2$, and the left and right limits of g'_U equal 0. Therefore, both g_U and g'_U are continuous at b .

For $p < 1/2$, the only discontinuity for g_U can possibly occur at $c = (1-2p)/(2-2p)$. However at this point $\phi(c) = 2c = (1-2p)/(1-p)$ so that $\frac{(1-p)(2-\phi(c))}{1-\phi(c)} = \frac{1-p}{p}$. Hence the left and right limits of g_U equal $(p^{-1}-1)^c/2$, and the left and right limits of g'_U equal $\log(p^{-1}-1)(p^{-1}-1)^c/2$. Therefore, both g_U and g'_U are continuous at c .

(c) Note that g_U is positive everywhere, so $\log(g_U)$ is (strictly) increasing if and only if g_U is (strictly) increasing. By the formulae for g'_U in part (b), it is easy to see that g'_U is always non-negative, and is positive when $g_U(a) > 1/2$. To show $\log(g_U)$ is convex, we need to show its derivative, i.e. g'_U/g_U is increasing. This also follows from part (b), noting that ϕ is increasing by Lemma 14(c). ■